Vol.2020-CG-180 No.7
Vol.2020-DCC-26 No.7
Vol.2020-CVIM-223 No.7
2020/11/5

インスタンスマップからの画像生成と主観による自然さの評価

大石 涼火[a)]　数藤 恭子[b)]

概要：本研究は，画像のピクセルごとのクラス分類情報が含まれている画像 (以下，セマンティック画像と呼ぶ) とインスタンス (車，人などの個々の物体) 情報が含まれている画像 (以下，インスタンス画像と呼ぶ) とを用いることにより，セマンティック画像から多様なインスタンスを含むリアルな画像を生成する GAN を提案する．Generator にインスタンスを区別する構造を持たせ，また，インスタンス内部の各画素に位置情報を持たせる位置マップを自動生成して同時に学習することで，(1) 空間的に連続して出 現する同一ラベル物体間の境界の明瞭化，(2) 画像生成におけるインスタンスの多様性向上，の両立を図る．(1)(2) の効果について，生成画像の質の定量評価及び，生成画像の自然さの主観評価を実施した．(1) の効果による FID の大幅な改善，(2) の効果による mIoU の改善，主観評価で提案手法が従来手法よりも自然な画像として選ばれやすいことを確認した．

キーワード：Deep Learning, 画像生成, 敵対的生成ネットワーク (GAN), Image Synthesis

Abstract: We propose a method to synthetic image that generates images with various instances from images with dense labels of object classes (semantic images) and images with instance labels of objects like car or person (instance images). Our network has a module to identify instances, and a module to generate position maps which has the pixel-wise position information inside of each instance in the image. These architectures enable us to realize satisfying both (1) Clarifying the boundaries between instances of the same label which appear consecutively in the same image, and (2)Increasing varieties of instances in generating images. The experimental results show that FID and mIoU are improved. The subjective evaluation also shows that the generated images by the proposed method tend to be selected as more natural images than a conventional method.

Keywords: Deep Learning, Image Generation, Generative Adversarial Network (GAN), Image Synthesis

1. はじめに

近年 GAN(Generativer Adversarial Networks)[1] による画像生成モデルの研究と計算機パワーの進展により，高解像度の画像を生成することが可能になりつつある．また，単に実写のように生成するだけでなく，属性を自在に制御する構造を取り入れた様々な派生系の GAN が提案されている [2][4][5]．本研究は，生成オブジェクトの属性制御において画像中に同一カテゴリのオブジェクトの複数のインスタンスが含まれる場合に，インスタンスの単位で属性制御を行うことを目的としている．GauGAN[11] をはじめとした従来の画像生成では，同一ラベル（例えば車，人など）の異なる物体を区別する情報が考慮されていないため，しばしば類似したインスタンスが頻出する．また，色やテクスチャが似通ってしまうことで物体間の境界が明確に生成できなくなったり，境界がなくなり複数の物体が大きな一つの物体になってしまうなどの問題がある．本研究では，学習データにインスタンス画像を追加するとともに，Generator にインスタンスを区別させるため，インスタンスレベルで特徴を区別する構造を導入する．また，インスタンスベースでの学習を安定させるため，物体領域の各画素にアンカーからの相対位置を埋め込んだ位置マップを導入する．これらの効果について定量評価および主観評価を行う．本手法での生成例を図 1 に示す．

2. 従来手法

従来の GAN による画像生成の課題の一つに生成物体の柔軟な属性制御がある．DCGAN[3] では Generator がラン

a) 6519001o@st.toho-u.jp
b) kyoko.sudo@sci.toho-u.ac.jp

Vol.2020-CG-180 No.7
Vol.2020-DCC-26 No.7
Vol.2020-CVIM-223 No.7
2020/11/5

図 1: インスタンスマップと位置マップを用いた提案手法による生成画像

ダムノイズを (C(チャネル数))$\times 4 \times 4$ などのテンソルに変換して画像生成を行うが, 高解像度の画像生成が困難であった．これに対し StyleGAN[10] では入力をランダムノイズではなく定数テンソル $C \times 4 \times 4$ とし，各解像度においてランダムノイズに全結合ネットワークを施し，潜在変数として入力する Adaptive Instance Normalization(AdaIN)[8] を用いる．これにより，各解像度での特徴を分離することが可能となる．例えば低解像度の特徴を固定し，高解像度の特徴のみを補間すると，画像の構造は維持されたまま色合いのみが変化することが示されている．Multimodal Unsupervised Image-to-image Translation(MUNIT)[9] では AdaIN を用いたスタイル変換を行うことにより，入力画像から複数出力のスタイル変換を行うことが可能となった．GauGAN[11] では，CNN を施したセマンティック画像のスタイル特徴を空間情報付き AdaIN パラメータとして扱う．また，実画像からの特徴ベクトルを入力とする際，その分布が標準正規分布に近づくように最適化することで，合成画像のセマンティック情報を維持したままスタイルを変更することが可能である．一方，車や人などをはじめとした各物体（インスタンス）は，実世界では様々な特徴をもっている．それと同じように，インスタンスのスタイルを制御するには，同一ラベルの異なる物体を区別して生成する必要がある．それには，インスタンス画像の利用が考えられる．しかし，GauGAN はセマンティック画像からの画像合成であり，固定長のチャネル数を想定した構造である．インスタンス画像はチャネル数が物体の個数と同じ可変長であることから GauGAN の手法をそのまま適用することはできない．そこで，可変長のインスタンス画像を固定長の特徴マップに埋め込む手法を提案する．

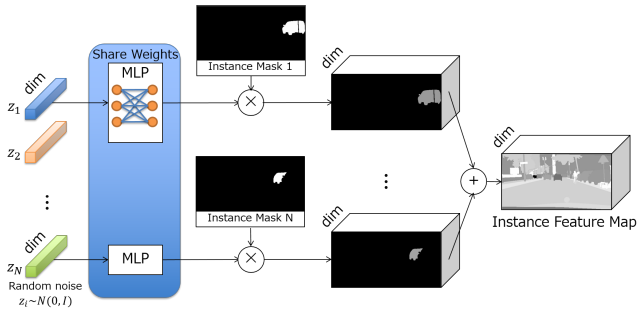

図 2: インスタンスマップを生成するネットワーク構成．各インスタンス領域全体にスタイルベクトルが埋め込まれている．それらの和をとることによりインスタンスの個数によらないスタイル特徴が埋め込まれた画像が獲得される．

3. 提案手法

3.1 Multi Instance MLP の提案

GauGAN[11] では，CNN を施したセマンティック画像のスタイル特徴を空間情報付き AdaIN パラメータとして扱う．それによって，セマンティック画像の同一ラベルには同じスタイル特徴が適用される．本研究では各可変長インスタンスに異なる特徴を与える Multi Instence MLP(MIMLP) を提案する．具体的には，全結合ネットワークからスタイルベクトルへの全結合ネットワーク写像（MLP）をマルチインスタンス（MI）に拡張する．提案手法の全体構成は図 2 のようになる．

あるインスタンス画像の 1 つの物体のマスクを x_{ins}，インスタンスの個数を N 個とする．まず i 番目のインスタンスについて正規分布からのベクトル $z^{(i)} \in \mathbb{R}^{ch}$ を入力とする．ここ特徴量 x をテンソルとし，$\mu(x), \sigma(x)$ をそれぞれ x の平均，標準偏差，$\varphi_\mu(z), \varphi_\sigma(z)$ を z からの特徴とすると正規化の式は以下で表される．これは GauGAN の SPADE[11] を改良したものである．正規化の構成は図 3 のようになる．

$$MIMLP(x, z^{(i)}, x_{ins}^{(i)}) = (\varphi_\sigma(z^{(i)}) \frac{x - \mu(x)}{\sigma(x)} + \varphi_\mu(z^{(i)})) x_{ins}^{(i)}. \quad (1)$$

式 (1) は $i = 1$ のとき，AdaIN と同様の働きを持つ．

3.2 境界情報のための位置マップの挿入

本研究では，インスタンスを分離する効果を狙い，特徴マップの物体領域の各画素にアンカーからの相対位置を埋め込んだ位置マップ (positional encodings) を導入する．位置マップの作成方法を図 4 で示す．この位置マップにより，隣接している物体の潜在変数が類似している状態であっても，境界情報が失われないことが期待される．位置マップのそれぞれの画素値は，時系列情報を正弦波に変換した positional encodings を埋め込み特徴に加える Ashish らの方法 [6] に倣い，インスタンス内部の位置情報を正弦波に変換する．位置マップのテンソルサイズは $C \times H \times W (C = 128)$ とし，i 番目のチャネルの座標 (x, y) における値を $PE(i, x, y)$ とする．各インスタンス領域に対

Vol.2020-CG-180 No.7
Vol.2020-DCC-26 No.7
Vol.2020-CVIM-223 No.7
2020/11/5

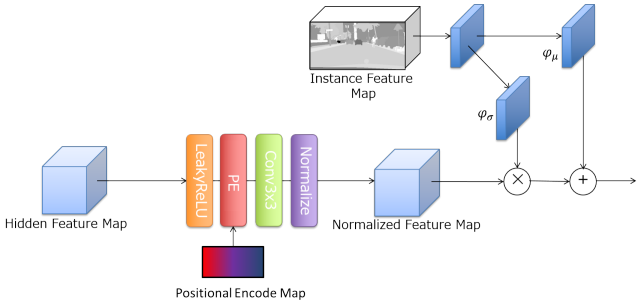

図 3: SPADE[11] の構造．MIMLP によって獲得されたスタイル特徴が埋め込まれた画像を AdaIN パラメータとし，各インスタンスごとに異なるパラメータで AdaIN を行う．

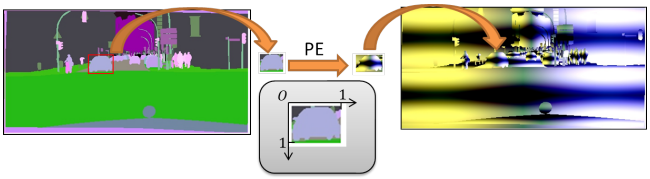

図 4: 位置マップの作成方法．インスタンスマップから各インスタンスに対して矩形領域を抽出し，矩形の左上，右下がそれぞれ $(0,0),(1,1)$ となるような画像を計算する．

し，インスタンスの矩形を (x_1, y_1, x_1+W, y_1+H) とすると，i 番目のチャネルの座標 (x,y) に挿入される位置マップの画素値 $PE(x,y)$ は以下の式で表される．

$$
\begin{aligned}
PE(4i+k,x,y) &= (p(i,x),p(i,y)) \odot d(k) \qquad (2)\\
p(i,x) &= sin((2\frac{x-x_1}{W}-1)\frac{4i}{C})\\
p(i,y) &= sin((2\frac{y-y_1}{H}-1)\frac{4i}{C})\\
d(k_{\in 0,1,2,3}) &= ((1,1),(-1,1),(1,-1),(-1,-1))
\end{aligned}
$$

4. 実験

4.1 データセットと実験概要

実験には，CityScapes Dataset[12][13] を用いる．CityScapes Dataset には，実画像，画素単位でラベルづけされたセマンティック画像，車や人など一部のカテゴリのインスタンスごとのセグメンテーションが施されたインスタンス画像が含まれる．なお，建物やフェンスなどのインスタンスごとのセグメンテーション対象でないラベルについては，本研究の多様なインスタンス生成の対象としない．モデルの評価は，GAN の評価に一般的に用いられる FID などの評価尺度による定量評価と，生成画像の質を目視で比較評価する主観評価の 2 種類を行う．以下セマンティックマップのラベルから生成を行う GauGAN（従来手法）を L，L にインスタンスマップを追加したもの（提案手法 1）を L+I，L+I に位置マップを追加したもの（提案手法 2）を L+I+P とする．各手法のモジュールの関係性をベン図で表したものを図 5 に示す．テストデータ 500 枚における生成結果を図 7，定量評価におけるスコアを表 1 に，主観評価の結果を表 2 にそれぞれ示す．

4.2 定量評価

定量評価の指標として独自の評価指標である MICV および，FID[7]，mIoU の 3 つを用いる．

MICV

インスタンスごとのバリエーションを定量評価する指標として，Mean Instance Color Variation (MICV) を提案する．最初にバリエーションを評価したいラベルを一つ選択する．（例として車，人など）次に生成画像から選択したラベル領域をインスタンスごとにマスキングを行う．その後，各インスタンスごとに RGB 値のそれぞれの平均値を求める．あるラベルのインスタンスが $n(\geqq 2)$ 個以上ある画像 x_i に対し，j 番目の対象ラベル領域の平均値 $\mu(x_i^{(j)})$ の標準偏差をその画像の Instance Color Variation(ICV) とする．すべての評価画像に対し ICV を導出，各画像の ICV に対してデータセット上 (N 枚) の平均値を MICV とする．したがって，MICV は式 (3) となる．MICV は必ず 0 以上の値をとり，数値が大きいほどインスタンスごとのバリエーションは豊富であると考えられる．MICV は以下の式で表される．MICV はインスタンスのカラーバリエーションを評価するために導入した指標である．MICV で選択するラベルは，ラベル情報とインスタンス情報がともに含まれている車と人を選択する．

$$
MICV = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_i-\mu(x_i^{(j)}))^2} \qquad (3)
$$

FID

Fréchet Inception Distance (FID)[7] による評価は画像生成分野などで用いられており，2 つの画像集合間における学習済み Inception モデルから得た特徴ベクトルとの距離を測るものである．FID が 0 に近いほど実画像と同程度の多様性を保つと考えられる．

mIoU

mIoU では，セマンティックセグメンテーションモデルの一つである学習済み PSPNet[14] を用いることで生成画像にセグメンテーションを行い，正解ラベル画像とのピクセルごとによる誤差を評価する．mIoU は 0 から 1 の間をとり，1 に近いほど良い評価となる．良い評価のときは生成画像が各ラベルを分離できるような画像を生成できていると考えられる．

4.3 主観評価

主観評価はアンケート形式で行う．同じシーン（同じ正解実画像）に対応する正解実画像，L(従来手法)，L+I(提案手法 1)，L+I+P(提案手法 2) の 4 種類の中から無作為に 2 種類を選び，各種類につき 1 枚ずつの画像を並べて提

Vol.2020-CG-180 No.7
Vol.2020-DCC-26 No.7
Vol.2020-CVIM-223 No.7
2020/11/5

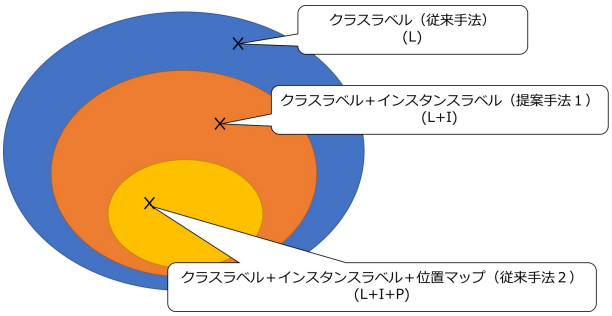

図 5: 提案手法 1 は従来手法にインスタンスラベルを追加したもの，提案手法 2 は提案手法 1 に位置マップを追加したものである．

図 6: 主観評価アンケートの調査画面．左右に並んでいる２枚の画像からより本物らしいものをクリックで選択する．選択後，次の問題に切り替わる．

示する (図 6)．被験者は表示された 2 枚の画像からより本物に近い方を選択（クリック）する．選択をする際は，(1) 画像が表示されてから 5 秒以内に画像をクリックする，(2) 時間無制限で画像をクリックする，の２種類のパターンでの実験を行う．(1) は，短時間での選択により「瞬間的」な印象で評価する狙いがある．なお，5 秒を過ぎた場合は回答されずに別の問題へと切り替わる．(2) は，時間制限を設けないため「注意深く」画像を確認できることから，インスタンスレベルでの認識力を調査できると考えられる．被験者は成人 16 名であり，(1)(2) 共に 100 枚ずつの選択を行なった．

5. 考察

5.1 定量評価

インスタンス画像の導入により mIoU，MICV が向上することを確認した．FID は位置マップの有無にはよらないがインスタンス画像を与えたことで，インスタンス画像を与えない場合と比べて大幅な改善がみられた．図 7 から，インスタンス画像の追加により，同一ラベルの個々の物体間において，それぞれ異なる特徴を持つような結果となっていることがわかる．また，位置マップの追加により，物体間の境界情報がより失われることなく生成が可能となることから，mIoU の精度が向上したと考えられる．位置マップの有無については，FID は同程度の精度であるが，MICV および mIoU の向上が確認された．位置マップ無しの場合，MIMLP に類似した乱数が入力されるとインスタンス情報が失われてしまう．位置マップありの場合は乱数によらずインスタンス情報が失われないため，精度向上したと考えられる．

表 1: 提案手法における位置マップの効果の評価．FID は小さいほど良い評価，FID 以外は大きいほど良い評価である．

Method	MICV ($\times 10^5$) Car	MICV ($\times 10^5$) Person	FID	mIoU
L+I+P(提案手法 2)	**5.12**	**6.06**	16.84	**0.502**
L+I(提案手法 1)	4.02	4.76	**16.26**	0.479
L(従来手法)	3.76	4.05	23.22	0.457

表 2: 主観評価におけるアンケート結果．従来手法であるクラスラベル L，従来手法である L にインスタンスマップ I を加えたもの L+I および，L+I に位置マップ P を加えたもの L+I+P とする．2 つのデータセット上から同じシーンの画像を選び，行に対応するデータセットが選ばれた割合を示している．

(1). 5 秒以内に選択.

データセット	L	L+I	L+I+P
L(従来手法)	-	0.44	0.31
L+I(提案手法 1)	**0.56**	-	0.4
L+I+P(提案手法 2)	**0.69**	**0.6**	-

(2). 時間無制限で選択.

データセット	L	L+I	L+I+P
L(従来手法)	-	0.40	0.34
L+I(提案手法 1)	**0.6**	-	0.43
L+I+P(提案手法 2)	**0.66**	**0.57**	-

5.2 主観評価

L(従来手法) と比較して L+I(提案手法 1), L+I+P(提案手法 2) ともに従来手法を上回る結果となった．(1) では，L+I+P(提案手法 2) と L(従来手法) を比較したところ，L+I+P(提案手法 2) が選ばれる割合が 69%となった．L(従来手法) で生成された画像は提案手法と比較して同一クラスの物体の境界が安定して生成されない傾向があることから，2 つの物体をあたかも 1 つの物体のように生成してしまい，色などが類似する．5 秒という短時間の中で本物を判断するために，色などで判断されたと考えられる．(2) では L(従来手法) と L+I(提案手法 1) を比較したところ，60%と (1) の 56%を上回る結果となった．このことから，時間無制限で選択を行うことにより同一クラスの隣接した物体のバリエーションをより判断できると考えられる．一方，L(従来手法) と L+I+P(提案手法 2) を比較したところ，提案手法 2 が優ってはいるものの (1) よりは低い 66%となった．これは，L+I+P(提案手法 2) では入力潜在変数に依存しない位置マップが加算されるため，畳み込み層が安定したエッジを検出できることから，同一クラスの隣接した物体のバリエーションが豊富でなくても安定した学習

ができるからであると考えられる．

6. まとめ

本研究では，GAN による画像生成において，同一ラベルの物体の多様なインスタンスの生成を可能にするネットワークを提案した．従来では CNN における局所特徴の影響でインスタンスが同じスタイル特徴をもち，複数の隣り合うインスタンスが繋がってしまう等の問題があったが，提案手法によりインスタンスの分離性能が向上することを示した．具体的には，セマンティック画像からの画像生成手法において，インスタンス画像と位置マップを導入したネットワーク構造により，同一ラベル物体のインスタンスを数によらず個別に属性パラメータの制御を可能にすることで，画像生成精度の向上と多様なインスタンス生成を可能にした．今後は道路以外のシーンなど，他のドメインのデータセットについても提案手法の有効性を確認する．本手法はインスタンス画像が利用可能であることを前提としているが，アノテーションのコストが高いことから，インスタンス画像はデータセットの一部にのみあればよい手法を検討する．

参考文献

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets.” In *Neural Information Processing Systems (NeurIPS)*, 2014.

[2] M. Mirza and S. Osindero. “Conditional generative adversarial nets.” *CoRR*, abs/1411.1784, 2014.

[3] A. Radford, L. Metz, and S. Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks.” arXiv preprint arXiv:1511.06434, 2015.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. “Image-to-image translation with conditional adversarial networks.” In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. “High-resolution image synthesis and semantic manipulation with conditional gans.” In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Ashish et al., “High-resolution image synthesis and semantic manipulation with conditional gans.” In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. “GANs trained by a two time-scale update rule converge to a local nash equilibrium.” In *Neural Information Processing Systems (NIPS)*, pp. 6629–6640, 2017.

[8] X. Huang and S. Belongie. “Arbitrary style transfer in real-time with adaptive instance normalization.” In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[9] X. Huang, M. Liu and S. Belongie and J. Kautz. “Multimodal Unsupervised Image-to-image Translation.” In *European Conference on Computer Vision (ECCV)*, 2018

[10] Karras et al., “A Style-Based Generator Architecture for Generative Adversarial Networks.” *CoRR*, abs/1812.04948, 2018.

[11] Park et al., “Semantic Image Synthesis with Spatially-Adaptive Normalization.” In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. “The cityscapes dataset for semantic urban scene understanding.” In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes Dataset.” In *CVPR Workshop on The Future of Datasets in Vision*, 2015.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. “Pyramid scene parsing network.” *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Vol.2020-CG-180 No.7
Vol.2020-DCC-26 No.7
Vol.2020-CVIM-223 No.7
2020/11/5

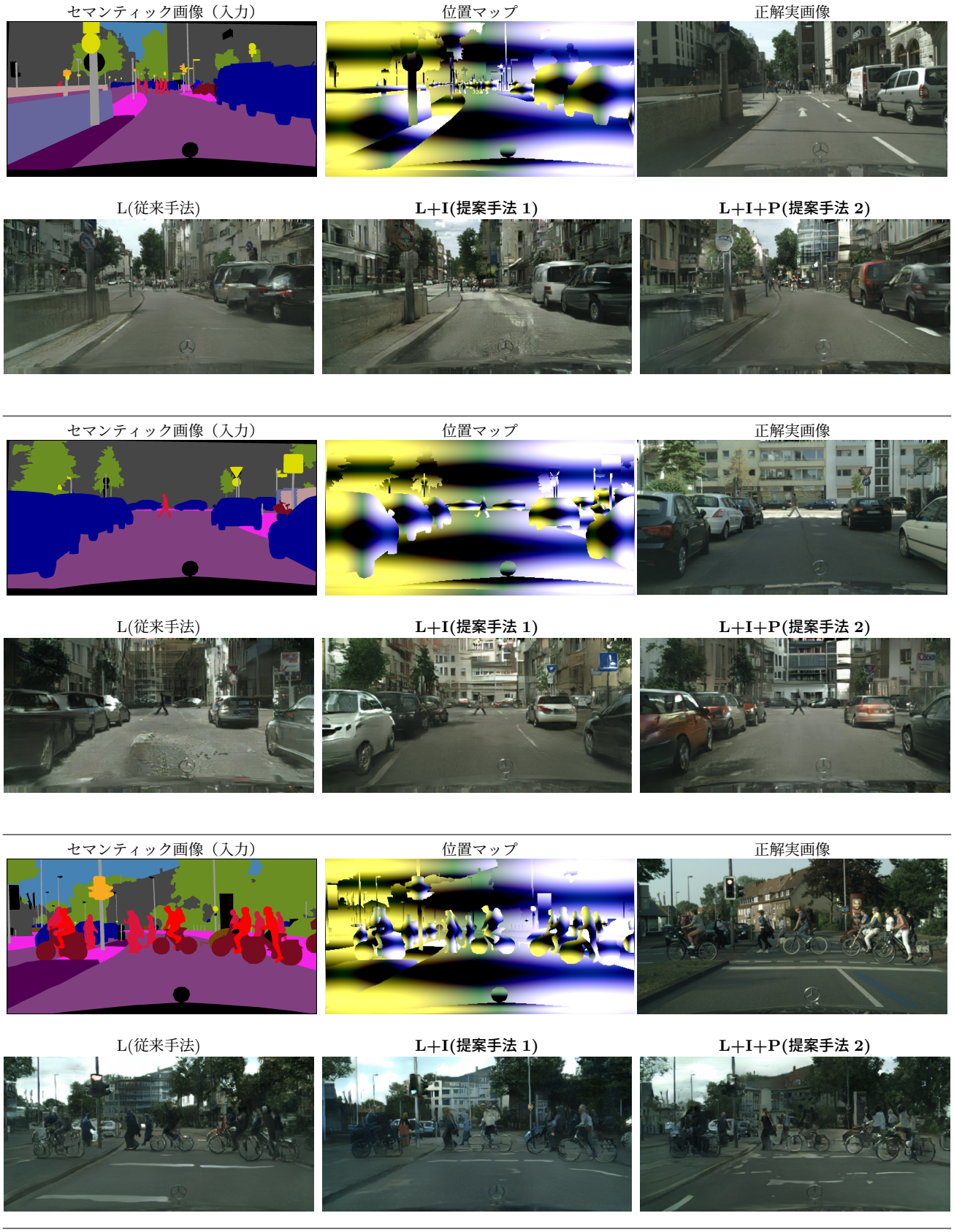

図 7: L(従来手法)，L+I(提案手法 1)，L+I+P(提案手法 2) の生成結果．インスタンス画像を導入しない場合，個々の物体の（インスタンス）特徴が類似したような結果となる．